FINAL REPORT
AUTOMATIC SPEAKER
RECOGNITION SYSTEM

CONTRACT N00014-84-C-2130

JULY 1984

PREPARED FOR:

NAVAL RESEARCH LABORATORY

WASHINGTON, D. C. 20375

Mr. Alan Higgins
Mr. Joe Naylor

ITT DEFENSE COMMUNICATIONS DIVISION
10060 CARROLL CANYON ROAD, SAN DIEGO 92131
619 578-3080

Table of Contents

1. INTRODUCTION

The Defense Communications Division of ITT (ITTDCD) has developed an automatic speaker recognition (ASR) system that meets the functional requirements defined in NRL's Statement of Work (RFP N 00014-84-R-M007). The Automatic Speaker Recognition System contract has two objectives, and the work is divided into two tasks. The objective of Task I is to evaluate the ASR system using a set of audio tapes supplied by NRL. The objective of Task II is to recommend a cost-effective means by which NRL can obtain an in-house facility for ASR research. Task I has been completed and the results of the system evaluation are given in this report.

The ASR system operates in real time using LPC-based features, is completely automatic in the sense that it requires no human interaction, and has a demonstrated capability of providing recognition accuracy in excess of 90 percent. The system is text independent, meaning that no restrictions are placed on the content of the speech used either for speaker modeling or recognition.

The GFM test material consists of eight audio tapes. Each tape contains 2 training phrases and 3 test phrases spoken by each of 20 speakers, 10 male and 10 female. One of the tapes contains clean speech recorded over a high-quality, wide-bandwidth channel. The remaining tapes are recordings of the same speech as the "clean" tape, after being passed through either a telephone simulator or one of six DoD voice processing systems.

This report is organized as follows. Chapter 2 is a short history of the development of the ASR system, both the algorithm and the implementation. Chapter 3 describes the methodology of the system testing, while Chapter 4 summarizes the test results. In Chapter 5, we discuss some further testing that was performed using the GFM test material. Conclusions derived from the contract work are given in Chapter 6. Finally, detailed results (statistics reported at half-second intervals) are given in Appendix A, which is bound as a separate volume. The data given in Appendix A are also recorded in ANSI standard format on the magnetic tape that accompanies this report.

2. ASR-SYSTEM BACKGROUND INFORMATION

2.1. Algorithm Development

The research and development leading to the current ITTDCD ASR system began with the *Automatic Speaker Recognition Comparison Study* contract (78-E275705-000). In this contract, the four leading text-independent speaker recognition techniques were compared using a common speech database to determine the most effective algorithm. The database used in the study was extremely large, containing eight and one half hours of conversational speech spoken by seventeen talkers. The speech had a high signal-to-noise ratio and was bandlimited to a telephone bandwidth of 300-3200 Hz. The four techniques compared were:

(1) The correlation of short and long term spectral averages as investigated by S. Pruzansky and M. V. Mathews[1].

(2) Cepstral measurements of long term spectral averages as investigated by S. Furui, F. Itakura, and S. Saito[2].

(3) Orthogonal linear prediction of the speech waveform as investigated by M. R. Sambur[3].

(4) Long term average LPC reflection coefficients, pitch, and overall gain of the speech waveform as investigated by J. D. Markel, B. T. Oshika, and A. H. Gray[4].

These four techniques differ in the type of parameters used for analysis, the classes of speech sounds used in the recognition, and the distance metric used in comparing the test sample with the speaker models. A more detailed description of the study and of the four methods is given in[5]. One result of the study was that the techniques that use LPC-derived parameters (Sambur's and Markel's techniques) performed significantly better on the text-dependent task than did the other two techniques.

The Automatic Speaker Recognition Comparison Study also provided a basis for the development by ITTDCD of an algorithm that improved upon Sambur's and Markel's techniques. This work was performed as part of the *Automatic Speaker and Language Recognition* (F30602-81-C-0134) contract with the Rome Air Development Center. This contract consisted of two

phases. The first was the algorithm development and comparison phase, in which various speaker recognition algorithms were implemented and tested to determine the optimal technique for satisfying the specific program requirements. The second phase of the program was to implement the selected technique in a real-time laboratory demonstration system, and to develop a convenient, easy-to-use operator interface to the system. The system performed text independent speaker recognition with very short utterances that simulated messages received over tactical military voice channels.

Two algorithms were developed and compared during the Automatic Speaker and Language contract. One of these, the Multi-Modal Model, is described in[6]. This algorithm performed very well with extremely short test utterances, but required relatively long training utterances, and was not robust in the presence of noise. The other algorithm, which proved to be the most promising, was known as the Multiple Parameter (MP) technique[7]. The MP technique is an enhancement of techniques developed by Markel and by Sambur. The technique involves calculating the mean, variance, and covariance of the reflection coefficients for all the frames in the segment of speech to be used as the model. This is done for each speaker to be identified. A mean vector for the reflection coefficients is calculated over the unknown speech, typically two to ten seconds long. This mean vector from the unknown is then compared to the mean vectors in the models using the Mahalanobis distance metric [4].

The MP technique is also based in part on work reported by Cheung et. al[8]. The strategy employed by Cheung corrects most of the problems found by other researchers in both Markel's technique and the Mahalanobis distance [9]. The strength of the MP technique is based on the combination of LPC-derived parameters, the use of a sub-band filter process, and adaptive thresholds for speech energy detection. This results in speech parameters which conform better to the assumptions that underlie the Mahalanobis distance.

During the course of the Speaker and Language Recognition contract, and in subsequent work, we tested the algorithm extensively. In addition, we further refined it to optimize its performance on data that has operational characteristics such as noise, level variation and channel

variations. We found that the MP technique performed well with short utterances and produced acceptable results for noisy speech after preprocessing by a speech enhancement unit. It also performed well when the testing and training material were transmitted over different channels. The robustness of the MP algorithm with respect to the above conditions made it the clear choice for text-independent speaker recognition.

2.2. Implementation

The current ITTDCD ASR system is a real-time laboratory demonstration system that uses the MP algorithm. It employs a flexible, easy-to-use operator interface. The system is currently implemented on a VAX-11/780 computer and a Floating Point Systems FPS-120B array processor. The computationally intensive portions of the algorithm, the LPC analysis, and the model distance calculations, are done in the array processor. The VAX is used as the system controller coordinating the interactions between the A/D, the D/A's, the FPS, and the operator interface.

Operation of the ITTDCD ASR system, or any other ASR system, is divided into two phases--a model generation phase, and a recognition phase. Model generation is an *off-line* process. Examples of a particular talker's speech are first digitized and the PCM is stored in a file. The model generation program is then run by specifying the input PCM file and the output model file name. Model generation proceeds with no operator intervention, and includes adaptive calculation of speech/silence thresholds.

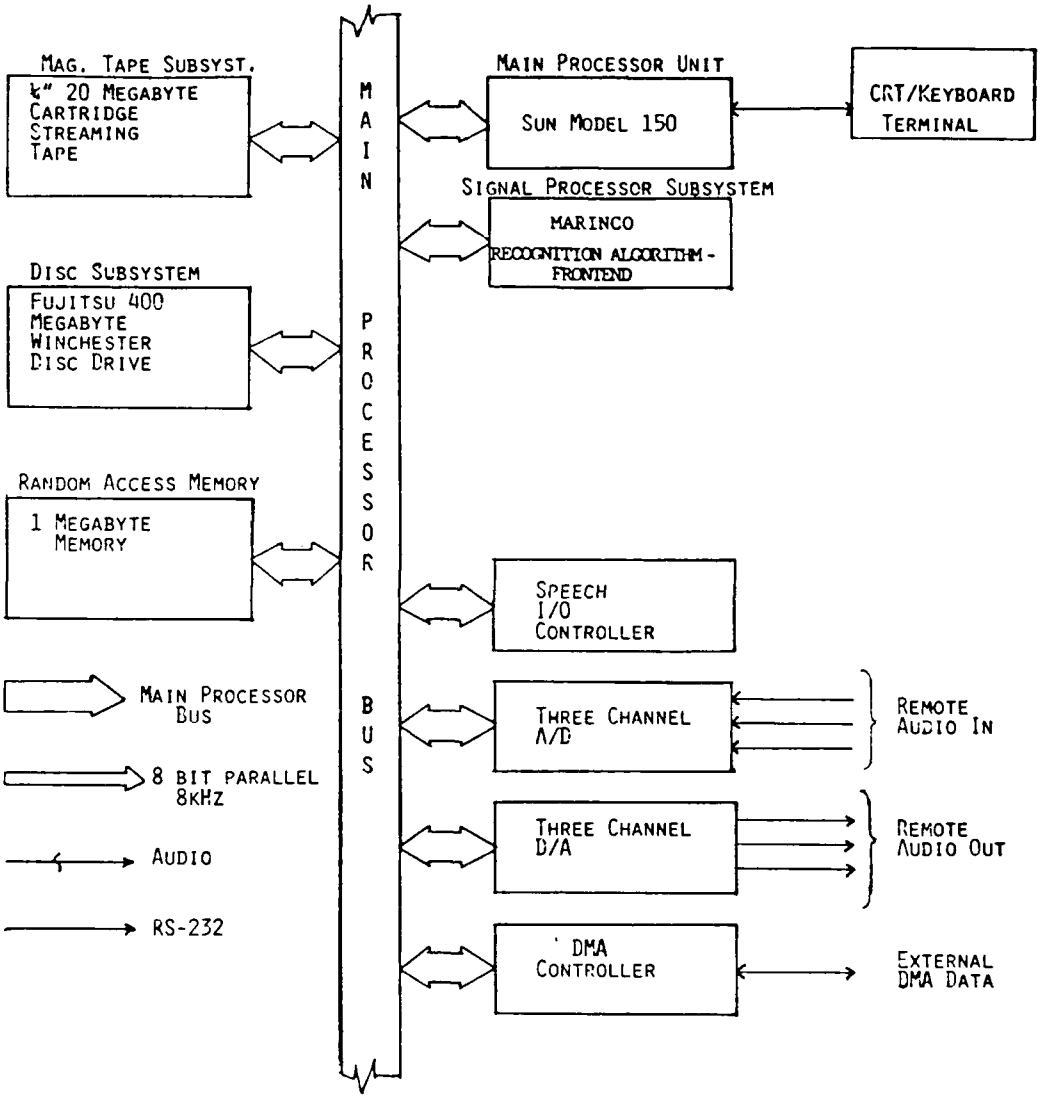
Once models are generated, the ASR system can be operated using either real-time analog input, or using previously-digitized PCM files. Recognition results are currently obtained each second and each recognition result is based on the previous N seconds of speech, where N is an operator-selectable parameter. To account for changes in talkers, if a silence region is found that exceeds 0.5 seconds, the recognition statistics are flushed from the system and the next recognition results that are output are based only on the speech that occurred since the silence interval. As in model generation, the threshold used to separate speech from silence is adaptively set by the recognition system, and no operator intervention is required.

2.3. Stand-Alone System

In addition to the real-time text-independent speaker recognition system described in Section 2.2, ITTDCD is presently building a low-cost stand-alone research system. The hardware block diagram for this design is shown in Figure 2.1. This system uses a Sun Work station as the main processor and user interface. The real-time signal processing and speaker-recognition algorithm is implemented in a small low-cost floating-point array processor. The system, as pictured has three channels of A/D and three D/A channels for data acquisition and playback. The entire system fits in a single equipment rack and uses a Berkeley Version of UNIX for the operating system.

The stand-alone ASR system will provide ITTDCD with a very valuable research tool in Speaker Recognition. During the last year, ITTDCD personnel analyzed many approaches to a stand alone real-time system implementation and found this to be the best low cost system for our needs. This experience and knowledge will aid us in determining the best low-cost solution for the Naval Research Laboratory.

Table 2.1: Hardware Block Diagram of ASR Stand-Alone System



3. TESTING METHODOLOGY

Figure 3.1 represents the format of the GFM audio tapes. There are a total of 100 different phrases (20 speakers times 2+3 phrases per speaker). Each phrase is a sentence of several seconds duration. A typical phrase is "Lift the square stone over the fence".

Each audio tape was processed as follows. The two or three phrases belonging to each speaker's training or test session were grouped together and digitized to form a single PCM file. There were therefore 40 such PCM files per tape. The PCM files derived from training sessions were used to generate speaker models, and the PCM files derived from test sessions were used in recognition trials.

Figure 3.2 is a block diagram of the testing methodology. The audio signal from the tape recorder was passed through a 4300-Hz anti-aliasing filter and into a 12-bit analog-to-digital (A/D) converter. The signal was sampled at 10000 samples per second. The A/D converter was turned on before the training or testing session started, and turned off after it ended. Therefore, several seconds of silence were included at the beginning and ending of the PCM record. These silence intervals were removed to reduce the amount of memory required to store the data. For this purpose, a simple energy-thresholding algorithm was used, and approximately one half second of "padding" was retained outside the detected endpoints to ensure that no speech sounds were cut off. The samples in each PCM file were then multiplied by a factor that was calculated to cause 1% of the samples to be clipped. The purpose of this step was to utilize the entire 12-bit range of the PCM representation.

One of the tests performed under the contract was a "Quantized LPC Parameters Test". This required a modification of the ASR system. The ASR algorithm (described in Appendix A of the proposal) uses 20 acoustic features derived from LPC analysis-- 10 reflection coefficients and 10 cepstral coefficients. LPC predictor coefficients and reflection coefficients are derived from the autocorrelation coefficients using Levinson's recursion. The predictor coefficients are then converted to cepstral coefficients. This processing is shown in Figure 3.3(a). For the quantized LPC parameters test, the system was modified as shown in Figure 3.3(b). The reflection

coefficients were quantized using the DoD standard quantization tables. The allocation of bits to quantization of each coefficient is shown in Table 3.1. The quantized reflection coefficients are converted first to predictor coefficients and finally to cepstral coefficients. The modification of Figure 3.3(b) applies to both model generation and recognition.

Coefficient:	1	2	3	4	5	6	7	8	9	10
Number of Bits:	5	5	5	5	4	4	4	4	3	2

Table 3.1: Bit Allocation for Reflection-Coefficient Quantization

The model-generation and recognition programs bandpass filter the input signal before computing the autocorrelation coefficients. As described in Appendix A of the proposal, this filtering is done in the frequency domain by windowing the power spectrum and inverse transforming. We have found that if the frequency-domain window does not have the same shape as the magnitude of a causal bandpass filter, then some frames may have unstable LPC parameters. This problem generally becomes more important for narrow bandwidth filters. In our initial testing, we used a simulated bandpass filter with a flat passband between 350 and 2500 Hz and with out-of-band rejection of 72 dB/octave. A small but significant number of the frames were unstable, as evidenced by reflection coefficients with magnitude greater than one. We therefore repeated the testing with a wider bandwidth filter. This second filter had a passband of 350-4000 Hz and the same band-edge slopes. No unstable frames were detected when this filter was used. The test results for both filters are given in the next section.

Figure 3.1: GFM Audio Tape Format

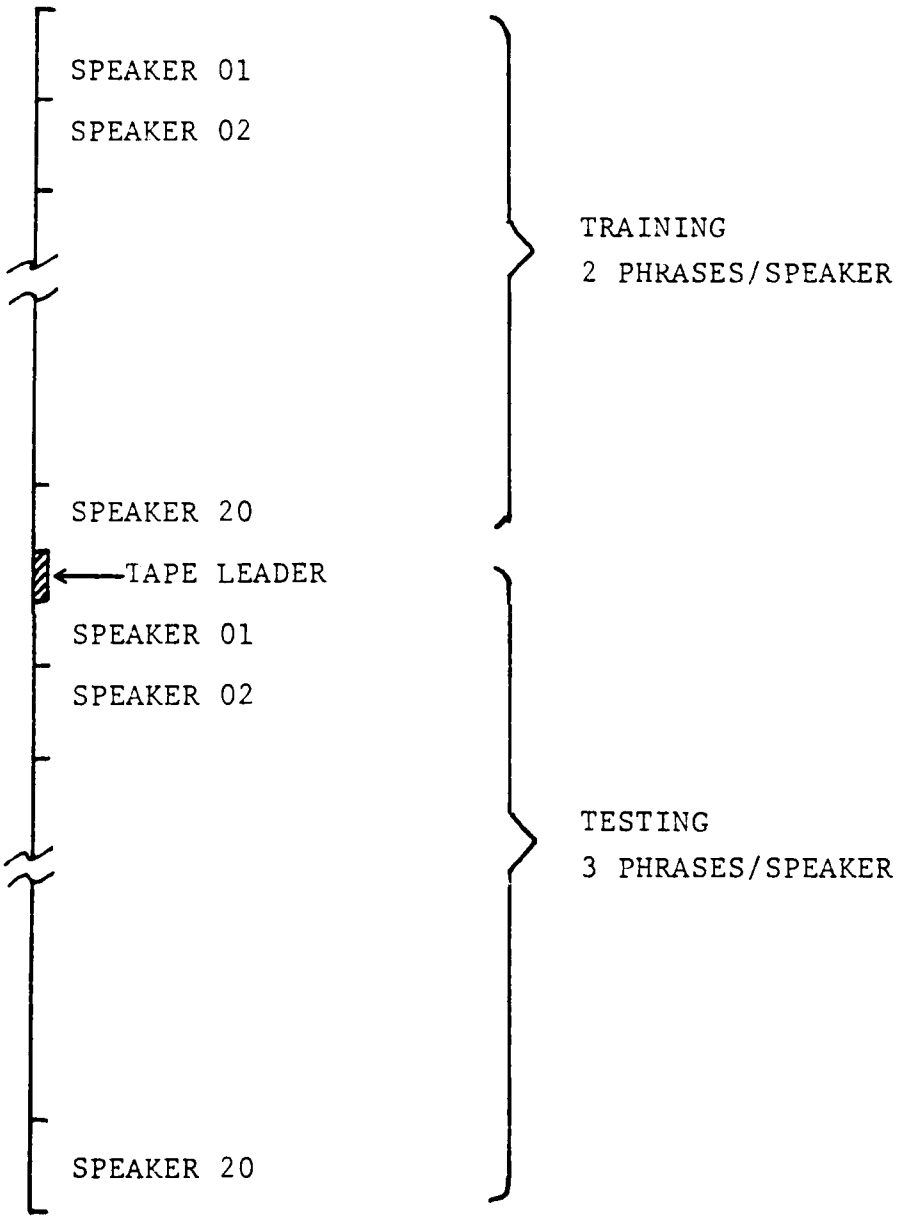


Figure 3.2: Block Diagram of Testing Methodology

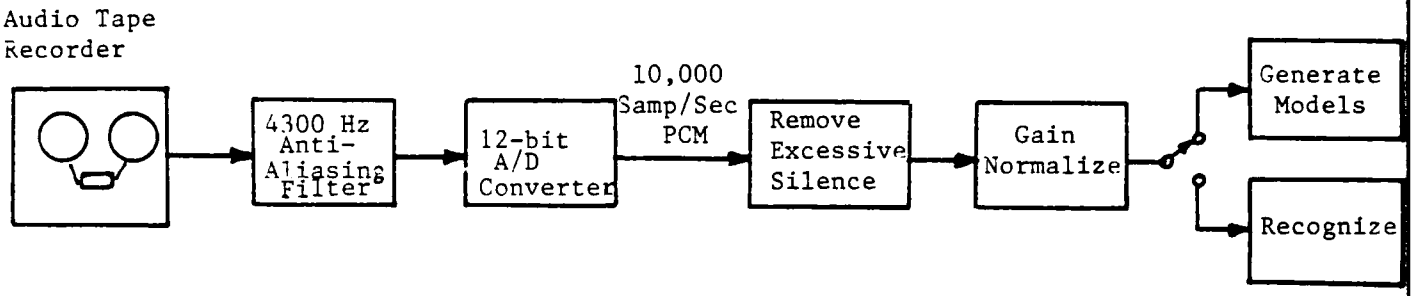
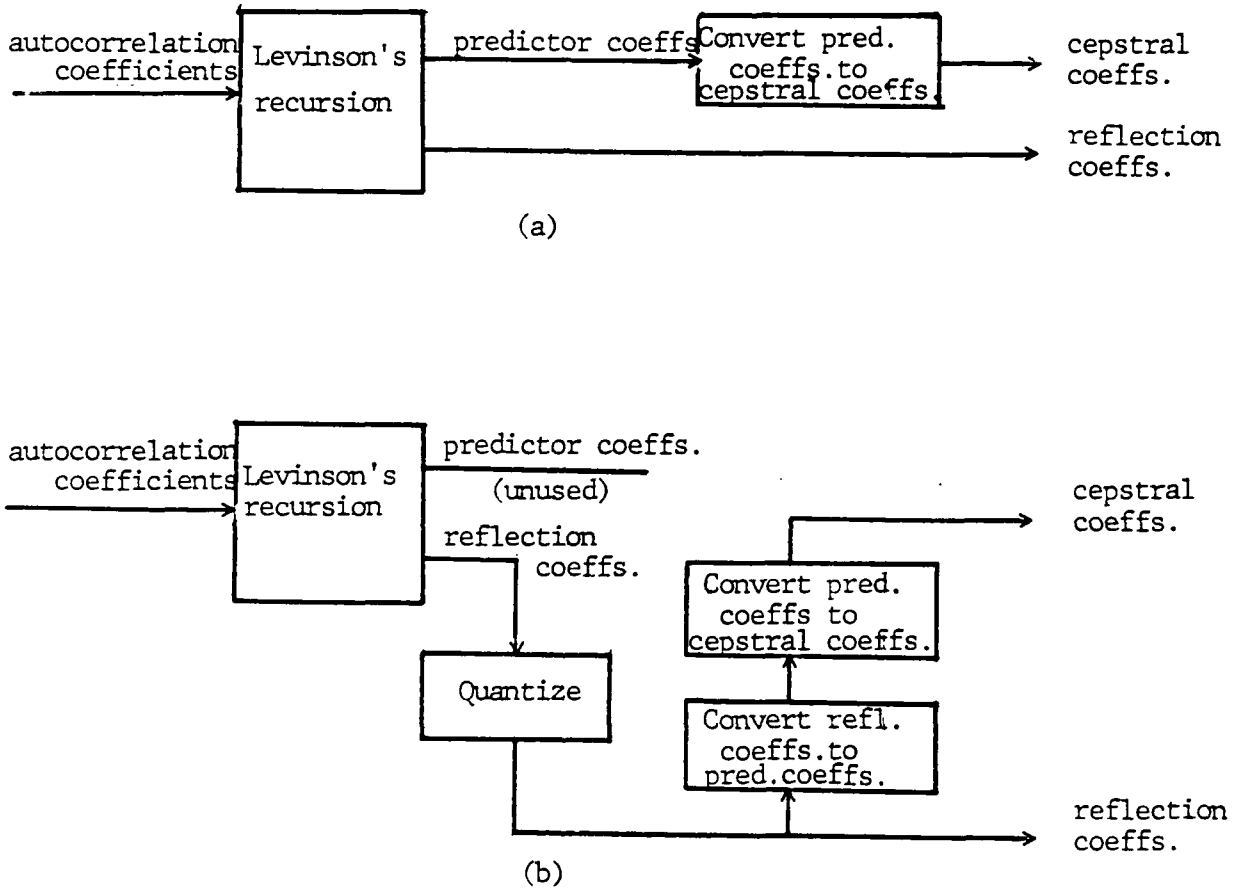


Figure 3.3: Signal Processing for Quantized LPC Parameters Test. (a) no quantization, (b) with quantization.



4. SUMMARY OF TEST RESULTS

Tables 4.1 and 4.2 summarize the results of the system testing, which was performed as described in the preceding chapter. Table 4.1 is for the 2500 Hz bandwidth, and Table 4.2 is for the 4000 Hz bandwidth. The test results are based upon all of the available speech data. Testing was performed only "within conditions". That is, speaker models and recognition trials were always derived from the same source tape. Each table has 9 columns, one for each source tape, and one for the quantized LPC parameters test (marked "Q"), which used the "clear" source tape. There are 20 rows, one for each speaker. A star indicates that the speaker was recognized correctly for the given condition. A number indicates a recognition error. The number given is the recognized speaker, while the row number is the actual speaker. The number of recognition errors is totalled at the bottom of the table, along with the recognition accuracy, measured with a granularity of 5%.

Referring to Table 4.1, Recognition accuracy for clear speech is 90%, as expected. Accuracy for the vocoded speech ranges from 70% to 95%. It is interesting that synthesized speech from two of the coders, v1 and v2, give better results than the original clear speech, although it is probable that the measured performance difference is not statistically significant. It is conceivable, however, that the analysis-synthesis procedures in question involve a quantization or smoothing that actually improves speaker discrimination. LPC parameter quantization alone reduces accuracy to 80%. By far the lowest accuracy (30%) was obtained for the telephone speech. We investigated this issue further, as discussed in the next chapter. One final note is that Speaker "04" was never recognized correctly, under any of the conditions. We have no explanation for this, however we verified that there were no procedural errors and that there was no unusual imbalance in the phonetic content of the training and testing phrases.

In Table 4.2, the range of scores is the same, but recognition accuracy for individual conditions changes by as much as 10%. Accuracy improves or remains the same under every condition except for coders v1 and v2, which degrade. Accuracy for clear speech and for coder v6 improve to 95%, while accuracy using quantized LPC parameters improves to 90%. Performance on the

telephone channel is unchanged.

It is interesting that recognition accuracy for coders v1 and v2 should degrade with the wider bandwidth, while accuracy for the other conditions improves or remains the same. A plausible explanation can be found in the type of signal processing performed in v1 and v2. These coders are Residual-Excited Linear Predictive (RELP) coders. They synthesize an accurate representation of the low-frequency portion, or baseband, of the speech signal and use spectrally-shaped noise in place of the high-frequency portion. Depending upon the method of high-frequency regeneration that is used in the RELP coder, the spectrum of the synthesized speech may differ significantly from that of the original speech in the high-frequency region. With this in mind, it is not surprising that v1 and v2 do not benefit from wider bandwidth. Speaker information is predominantly carried in the baseband, and the amount of additional speaker information carried in the high frequencies is outweighed by the inaccuracy of the spectral representation. Coder v2 degrades less than v1 because it probably has a wider baseband. Table 4.3 lists the coding techniques used in each of the six coders.

Speaker Number	Condition								
	cln	tel	v1	v2	v3	v4	v5	v6	Q
1	*	*	*	*	*	*	*	*	10
2	*	*	*	*	1	1	1	*	*
3	*	13	*	*	*	*	*	*	*
4	5	18	10	5	6	6	5	5	10
5	*	*	*	*	*	13	10	*	10
6	*	9	*	*	*	*	*	*	*
7	*	18	*	*	*	*	*	*	*
8	*	10	*	*	14	14	*	*	*
9	*	*	*	*	*	*	*	*	*
10	*	13	*	*	*	*	*	*	*
11	*	*	*	*	*	19	*	*	20
12	14	7	*	*	14	14	14	*	*
13	*	16	*	*	16	*	*	*	*
14	*	19	*	*	*	*	*	*	*
15	*	*	*	*	*	*	*	*	*
16	*	5	*	*	*	*	*	*	*
17	*	*	*	*	19	*	*	*	*
18	*	*	*	*	*	*	*	*	*
19	*	7	*	*	*	*	*	*	*
20	*	19	*	*	*	*	*	16	*
# Errors	2	14	1	1	6	6	4	2	4
% Recog	90	30	95	95	70	70	80	90	80

Table 4.1: Summary of Test Results for 2500-Hz Bandwidth

Speaker Number	Condition								
	cln	tel	v1	v2	v3	v4	v5	v6	Q
1	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	1	*	*
3	*	13	*	*	*	*	*	*	*
4	*	18	5	5	6	5	5	5	*
5	10	6	10	*	*	*	10	*	6
6	*	3	*	*	*	*	*	*	*
7	*	18	*	*	*	*	*	*	*
8	*	19	*	*	14	14	*	*	*
9	*	*	*	*	*	*	*	*	*
10	*	18	*	*	*	*	*	*	*
11	*	*	*	*	*	14	*	*	*
12	*	7	*	*	14	14	14	*	*
13	*	19	19	20	19	*	*	*	20
14	*	19	*	*	*	*	*	*	*
15	*	*	*	*	*	*	*	*	*
16	*	20	*	*	*	*	*	*	*
17	*	*	*	*	*	*	*	*	*
18	*	7	*	*	*	*	*	*	*
19	*	7	*	*	*	*	*	*	*
20	*	7	*	*	*	19	*	*	*
# Errors	1	14	3	2	4	5	4	1	2
% Recog	95	30	85	90	80	75	80	95	90

Table 4.2: Summary of Test Results for 4000-Hz Bandwidth

Coder Designation	Coding Method	Data Rate (bits/sec)
v1	Residual-Excited Linear Prediction (RELP)	9600
v2	RELP	16000
v3	Continuously-Variable Slope Delta Modulation (CVSD)	16000
v4	Adaptive Predictive Coding (APC)	9600
v5	Linear Predictive Coding (LPC)	2400
v6	PCM	64000

Table 4.3: Coding Techniques used in the ASR System Testing

5. FURTHER TESTING

We performed some further testing to answer two questions: (1) Why is accuracy for the telephone channel so low?, and (2) How sensitive are the results to changes in the threshold for frame selection?

5.1. Analysis of Telephone-Channel Results

As a first attempt to explain the poor performance on telephone data, we ran an experiment in which the training data itself was used for recognition. This resulted in perfect recognition, as it should, confirming that some differences existed between the speaker models. We then computed the long-term spectrum of the telephone data, averaged across the training utterances of all speakers. This spectrum is shown in Figure 5.1. Note that the bandwidth of the telephone channel is very limited, dropping 30 dB from its peak value at less than 2000 Hz. This channel is much worse than typical U. S. telephone lines and more closely resembles the poorer telephone lines of some European countries. For comparison, Figure 5.2 shows the frequency response of actual telephone lines measured during 5 separate calls from Baltimore to San Diego. The high-frequency band edge is at about 3300 Hz.

Figure 5.1: Long-Term Spectrum of Telephone Data

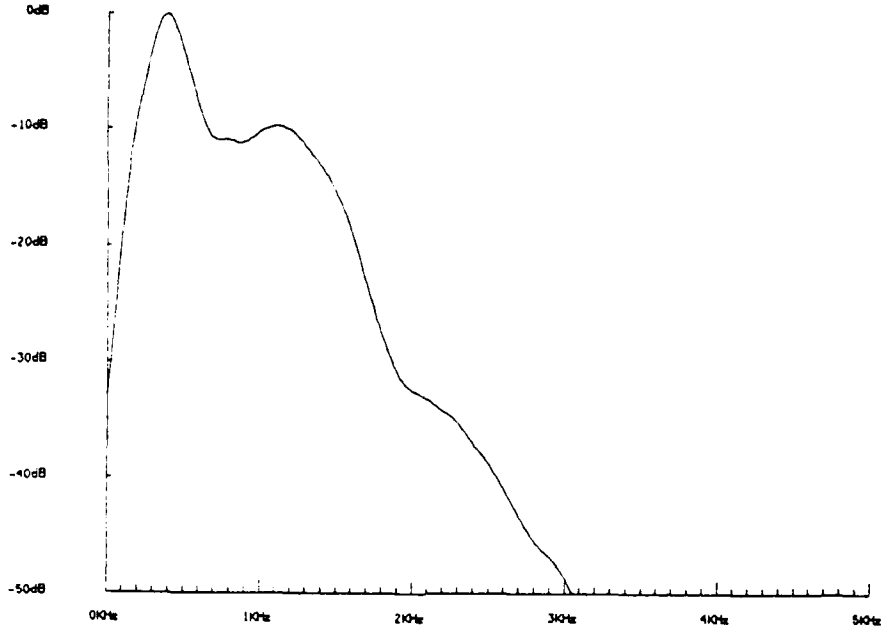
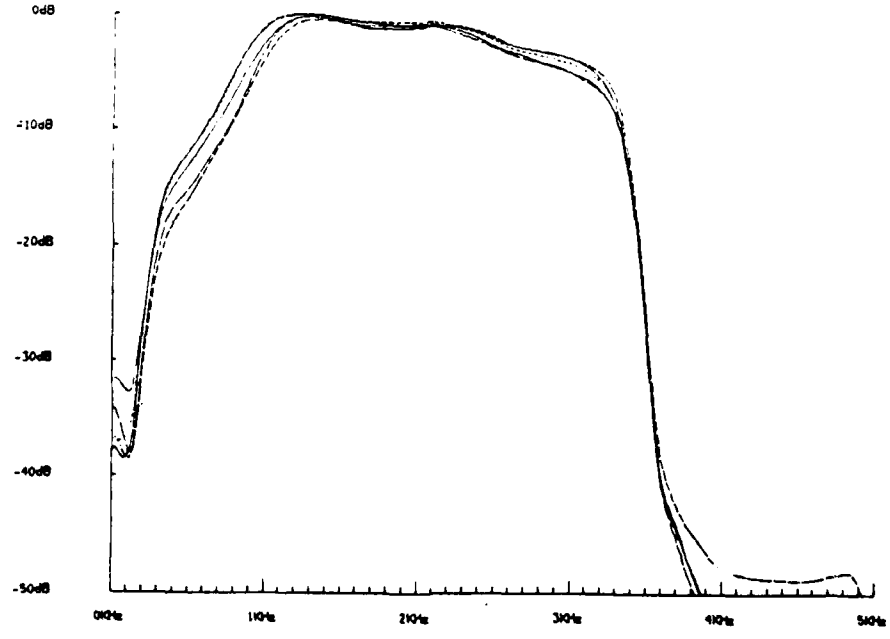


Figure 5.2: Measured Frequency Response of Five Long-Distance Telephone Lines



We suspected that the narrow bandwidth of the telephone channel was the cause of the performance degradation. We therefore ran a series of experiments in which the bandwidth of the clear-speech data used for recognition was successively reduced. This was done by adjusting the high-frequency band edge of the bandpass filter used in the model-generation and recognition programs. The first row of Table 5.1 shows the results of these experiments. Percentage recognition accuracy is given as a function of bandwidth. Accuracy is roughly constant for bandwidths of at least 2500 Hz, and drops rapidly for narrower bandwidths.

	Channel Bandwidth (Hz)						
	1750	2000	2250	2500	3000	3500	4000
Original Algorithm	5	10	5	90	85	90	95
Modified LPC	70	60	65	85	85	85	-

Table 5.1: Recognition Accuracy vs. Bandwidth

The extremely poor performance for narrow bandwidths was caused by ill-conditioning of the LPC computation. Because of the lack of energy at high frequencies, the spectral dynamic range of the bandpass filtered signal was very large. A simple modification of the algorithm was found to solve this problem. We multiplyed the energy in each frame (the 0th autocorrelation coefficient) by 1.0001 before computing the LPC coefficients. This effectively limits the spectral dynamic range by adding equal energy at all frequencies. Recognition accuracy for the modified algorithm is shown in the second row of Table 5.1.

Turning back to the telephone speech, we found that the recognition accuracy for the telephone tapes could be increased to 65% using a combination of the modified LPC computation and reduction of the high-frequency edge of the ASR bandpass filter to 1850 Hz., or roughly the band edge of the telephone channel. The later change reduces the influence of variations in the low-energy portion of the spectrum.

5.2. Investigation of Frame Selection

The ASR algorithm selects high-energy frames to use for speaker modeling and recognition.

To do this, it establishes an adaptive energy threshold by computing the minimum of a running histogram of the frame energies. The ASR system user may adjust the threshold in terms of the level in dB above or below the histogram minimum. We used a default level in the baseline testing of 3 dB below the minimum.

Since the frame-selection algorithm is adaptive, it requires a short time interval to converge upon a stable threshold at the beginning of each training or test utterance. However, these utterances are relatively short to begin with, so frames should not be wasted in the startup condition. We tested the system with the threshold set at 100 dB below the histogram minimum, thereby retaining all frames. The recognition accuracy was uniformly lower under every condition, with 75% accuracy on the clear speech. We repeated the experiment on clear speech with the threshold set at 3dB above the minimum, and found the performance to be unchanged relative to the baseline result. This indicates that the system performance is relatively insensitive to the threshold value, as long as it is in the vicinity of the histogram minimum, but that some frame selection is necessary to exclude silence frames.

6. CONCLUSION

We have found that the performance of the ASR system on clear speech is as expected, based on the length of the training and test phrases. The performance is robust with respect to transmission over DoD voice coders, and with respect to quantization of the LPC parameters. Performance does degrade sharply when the speech is transmitted over an extremely narrow bandwidth telephone channel, such as the one used in recording the GFM telephone tape. Based on measurements of actual long-distance U.S. telephone lines, we believe that the system would perform better in real use within the United States than this test indicates.

We are continuing to study and improve the ASR system. Our ongoing work in this area focuses both on the algorithm and implementation aspects of the system. We hope to continue to work with NRL in the future development of ASR systems. For the Automatic Speaker Recognition System contract, we shall now switch our effort to Task II, the System Implementation Study. At the end of this effort, in about 3 months, we shall deliver our recommendations in the System Implementation Report.

References

1. S. Pruzansky and M. V. Mathews, "Talker-recognition procedure based on analysis of variance," *J. Acoust. Soc. Amer.*, vol. 36, pp. 2041-2047, 1964.

2. S. Furui, F. Itakura, and S. Saito, "Talker recognition by long-time averaged speech spectrum," *Electron. Commun. Japan*, vol. 55-A, pp. 54-61, 1972.

3. M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-244, pp. 283-289, 1976.

4. J. D. Markel, B. T. Oshika, and A. H. Gray Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, pp. 330-337, 1977.

5. R. Wohlford, E. H. Wrench, and B. P. Landell, "A comparison of four techniques for automatic speaker recognition," *Proc. IEEE Internat. Conf. Acoust., Speech and Signal Process.*, no. 3, pp. 908-911, 1980.

6. K. P. Li and E. H. Wrench, Jr., "An approach to text-independent speaker recognition with short utterances," *Proc. IEEE Internat. Conf. Speech, Acoustic, Signal Process.*, no. to appear, 1983.

7. E. Wrench, *Final Report Automatic Speaker and Language Identification ContractF30602-81-C-0134*, RomeAir Development Center, 1983.

8. R. S. Cheung and B. A. Eisenstein, "Feature selection via dynamic programming for text-independent speaker identification," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-26, pp. 397-403, 1978.

9. R. Schwartz, S. Roucos, and M. Berouti, "The application of probability density estimation to text-independent speaker identification," *Proc. IEEE Internat. Conf. Acoust., Speech and Signal Process.*, pp. 1649-1652, 1982.